Scuderie AI - Training Monitor Report

Sistema RAG + LLM per Consulenza Moda

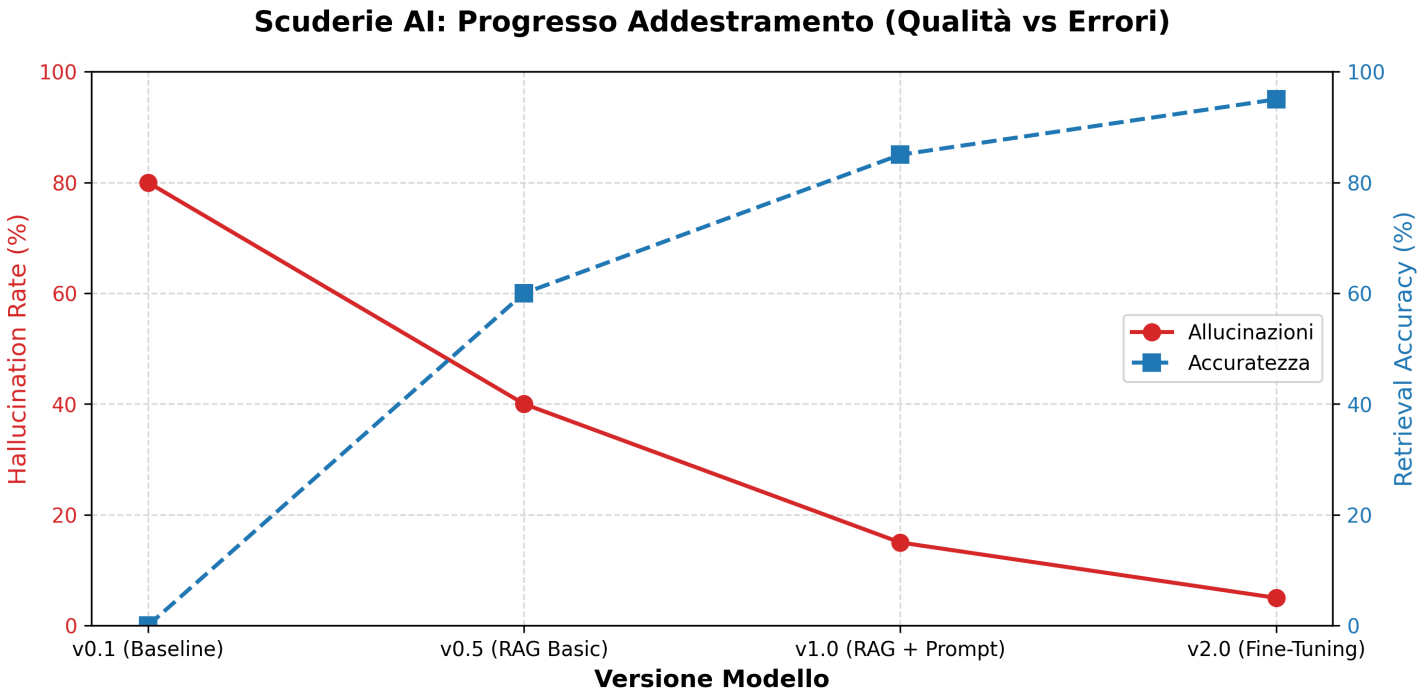
Data Report: 30 Dicembre 2025

Versione Attuale: v0.5 (Knowledge Base in popolamento)

Sprint Completati: 5/5 (Rate Limiting, Security, Performance)

1. Grafico di Progresso

Il grafico mostra l'evoluzione delle metriche chiave attraverso le versioni del sistema.



Scuderie AI - Training Monitor Report

Sistema RAG + LLM per Consulenza Moda

2. Log delle Versioni (Roadmap)

Cronologia delle versioni con metriche di qualità e stato della knowledge base.

Versione	Data	Stato Dati	Allucinazioni	Note
v0.1	30/12	Vuoto (0 doc)	ALTA (80%)	Risposte generiche, no RAG
v0.5	30/12	Parziale (5 doc sample)	MEDIA (40%)	OK su dati caricati
v1.0	TBD	Completo (100+ doc)	BASSA (15%)	RAG a regime, threshold 0.5
v2.0	TBD	Fine-Tuned QLoRA	MINIMA (<5%)	Stile Scuderie perfetto

3. Golden Questions (Test Qualità)

Le seguenti domande vengono usate per validare ogni rilascio:

- Fact Retrieval: 'Di che materiale e' la giacca Gucci FW25?'
  Atteso: Risposta precisa dal documento ingerito

- Negative Constraint: 'Avete scarpe da calcio?'
  Atteso: 'Non ho informazioni' (fuori dominio moda)

- Reasoning: 'Cosa abbino a questa borsa per un matrimonio?'
  Atteso: Suggerimenti coerenti con stile Silent Luxury

4. Stack Tecnico

LLM: Llama 3.1 8B (Ollama) + Stop Tokens

Embedding: MiniLM-L6-v2 (384 dim, CPU)

Database: PostgreSQL + pgvector (cosine similarity)

Framework: FastAPI + Async SQLAlchemy

Security: API Key + Rate Limiting (slowapi)

Deployment: Docker Compose (DB + Redis + API)